Data Collection and Preprocessing Phase

Date	9th July 2024
Team ID	SWTID1719999219
Project Title	Crystal Clear Vision: Revolutionizing Cataract Prediction through Transfer Learning Mastery
Maximum Marks	2 Marks

Data Quality Report

The Data Quality Report will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Retina Dataset Github	Data not split into train, validation and test directories.	High	Using os and shutil libraries. The os library will be used to create new directories for train, validation and test. The shutil library will be used to copy the files from the source folder to their destination train, validation and test folders.

Retina dataset from Github	The dataset has directories for normal, cataract, glaucoma and retinal disease. However, we require only normal and cataract directories.	Moderate	Using the 'rm' Shell command to remove unnecessary directories.
Retina dataset from Github	Image dimensions are too large.	Moderate	Images will have to be resized to a dimension of 224x224.
Retina dataset from Github	There are few augmented images.	Moderate	Data augmentation has to be applied.